Learning-based Spectrum Sharing and Spatial Reuse in mm-Wave Ultra Dense Networks

Chaoqiong Fan, Bin Li, Chenglin Zhao, Weisi Guo, Ying-Chang Liang, *Fellow, IEEE*

***Abstract*—In this paper, the throughput maximization of millimeter-wave (mm-Wave) ultra-dense networks (UDN) using dynamic spectrum sharing (DSS) is considered. Most of the existing works only allow temporal-domain access and admit at most one user at each time slot, resulting in significant under-utilization of spectrum resource, which will be less attractive to mm-wave UDN applications. A generalized temporal-spatial sharing scheme is proposed in this paper for UDN by exploiting the location information of incumbent devices, where multiple users are allowed to access each channel simultaneously via spatial separations. For distributed applications, the global information exchange among secondary users (SU) tends to be impractical, given the unaffordable signaling overhead and latency. Thus, a non-cooperative game with fine-grained two-dimensional reuse is formulated, which leads to a more efficient access strategy. It is then proved to be an exact potential game (EPG), which has at least one pure strategy Nash equilibrium (NE). Finally, an improved decentralized reinforcement learning algorithm is designed, with which SUs can learn from wireless environments and adapt towards to a NE point, relying on the individual observation and the historical action-reward (rather than the global information exchanging). The convergence efficiency of the new scheme is also rigorously proved. Numerical simulations are provided to validate the performances of the proposed schemes.**

***Index Terms*—Ultra-dense networks, millimeter-wave, temporal-spatial reuse, Nash equilibrium, decentralized learning**

I. Introduction

ATTRIBUTED to the explosive development of wireless communications, the data traffic has been growing in an exponential manner [1]. As one of the core features in the emerging 5G communications, ultra-dense networks (UDN) has attached extensive investigations [2]. In such a circumstance, a larger number of small cells with outnumbering users will be deployed crowdedly [2]–[4], which offers new features to wireless coverage, i.e., any given user would be in a very close distance to many small cells [5]. As a result, UDN remains strikingly different from traditional wireless networks. One major advantage of UDN is that, its coverage area becomes small enough to have a high probability of line-of-sight (LoS) transmission [6], [7] and, therefore, the low-power small-cell will gain prominence [5]. On the other hand, in such a context traditional techniques suffer from the poor quality of service (QoS), and new network paradigms to accommodate massive wireless devices will be critically needed [3].

Despite the potential of admitting the ever-growing devices, how to effectively control the network interference is of significance to UDN [8]. First, serious cross-link interference is inevitable among neighboring cells in the UDN proximity scenarios, which greatly degrades the performance [9]. To this end, deploying UDN in millimeter-wave (mm-Wave) band seems to be extremely attractive [10], due to a low risk of interference (i.e. the significant path-loss) and plenty of spectrum resources, as far as the low-power nature of small cell is further concerned. Second, the low spectrum efficiency limits the number of active devices and the overall network capacity [5]. It is recognized that dynamic spectrum sharing (DSS) is one effective approach to cope with the aforementioned challenges [11], [12]. By promoting the spectrum utilization via opportunistic spectrum access [13] and coordinating interference via listen-before-talk (LBT) techniques [14], DSS is of great promise to mitigating the interference [15], [16] and maximizing the network throughput [17]–[19].

There are plenty of studies on interference-aware dynamic access in the context of cognitive radio networks (CRN) [17]–[25]. In general, the shared users, or secondary users (SUs), are assumed to be capable of exchanging information of available spectrum resources, and can negotiate with each other according to various requirements. In [17], the resource allocation scheme for cognitive small-cell networks (C-SCNs) based on the cooperative bargaining game (CBG) is proposed. With the purpose of maximizing throughput and minimizing collisions, opportunistic spectrum access (OSA) premised on a local interaction game, which includes local altruistic game (LAG) and local congestion game (LCG), is studied [18]. A non-cooperative dynamic game for DSS is presented in [19]. In [20], the authors investigate a problem of aggregated interference from multiple SUs to primary user (PU). Other distributed schemes for sharing the sensing results are studied by [21]. In [22], self-organized spectrum access is used to mitigate the interference of SCN; while in [23], a stochastic learning approach based on the potential games is formulated to maximize the throughput of dynamic environments. In [24], a competition versus cooperation game on multiple-input single-output (MISO) interference channel is designed. A method of coordinated multi-users spectrum sharing in distributed antenna based CR system is presented in [25].

For UDN small-cells, existing schemes become less attractive. First, most schemes rely on an ideal assumption that SUs

Chaoqiong Fan, Bin Li and Chenglin Zhao are with the School of Information and Communication Engineering (SICE), Beijing University of Posts and Telecommunications (BUPT), Beijing,100876, China. (Email: stonebupt@gmail.com).

Weisi Guo is with the School of Engineering, University of Warwick, West Midlands, CV4 7AL, UK (E-mail: weisi.guo@warwick.ac.uk)

Ying-Chang Liang is with School of Electrical and Information Engineering, the University of Sydney, NSW, Australia, and with Institute for Infocomm Research, A*STAR, Singapore (email: liangyc@ieee.org).

have the full/global knowledge on: (1) the wireless environment, and (2) the complete (or semi-complete) information on actions taken by other partakers. Such assumptions will be unfortunately infeasible for UDN applications, as acquiring and exchanging such information will be resource-demanding (in terms of consumed time, power or bandwidth) and may lead to heavy signaling overheads as well as unaffordable latency. Second, probably due to the absence of PU's location information, most studies deal with the temporal-domain interference avoidance, i.e. only one SU is allowed to access vacant spectrum at one time slot. As far as the UDN is concerned, such DSS schemes with the simple sharing strategies will become inadequate, and more importantly, both the accommodated devices and the network capacity will be restricted. To the best of our knowledge, there are few works reported on the multiple-dimensional spectrum reuse (i.e. temporal and spatial) in the context of UDN.

In this study, we focus on the spectrum access and interference coordination in UDN, where there arises some new formidable challenges that most existing schemes fail to cope with. To be specific, how to realize the distributed spectrum access in UDN, and simultaneously, coordinate the intensive mutual interference via the affordable signaling overhead remains still as a key obstacle. In contrast to previous studies on overlay/underlay sharing, in this paper, we suggest a fine-grained and multiple-dimensional access scheme for UDN applications. We assume the partial information on PUs' locations will be available at SU, e.g., owing to the recently proposed deep sensing (DS) framework [26], [27], whilst the information of other SUs remains unknown. We are in particular interested in shared access with spatial reuse, where each SU can only be aware of its own channel selections and access reward. That is to say, each single SU link only perceives the signal-to-interference-and-noise ratio (SINR) of its receiver via some limited feedback. We then introduce a game-theoretic approach to identify the optimal accessing strategy. We employ the expectation of accumulated capacity as a utility function [25], and prove it is an exact potential game (EPG) which thereby has at least one pure strategy Nash equilibrium (NE) point. For the formulated game with only partial information, we further suggest a decentralized Q-learning algorithm, with which SUs learn from the individual action-reward history and adapt their behaviors towards to a NE point. To sum up, the main contributions are listed as follows:

1) For mm-Wave UDN, we established a new DSS model enabling temporal-spatial spectrum reuse, where multiple shared links are allowed to access the same channel at the same slot. An SINR temperature limit is specified to characterize the interference tolerance of shared receivers (i.e. a SU link below this limit will not access the channel at current slot). By configuring spatial beams flexibly, SUs aim to maximize its own transmission efficiency whilst restricting the interference to other receivers. Thus, this scheme achieves two-dimensional (i.e. temporal and spatial) spectrum sharing, which, by coordinating the cross-link interference, can accommodate more devices and further improve the network capacity.
2) Due to the more complicated problems and the higher performance requirements of the considered UDN scenario, directly applying existing game methods seem to be a profitless exercise. To combat this, it is urgent to redesign and reformulate the game model for UDN application. In this paper, we present a unified non-cooperative game framework for the DSS problem in distributed UDN scenarios. To make it more suitable for our considered complex scenario, we employ the expectation of SUs' reward as a utility function and the accumulated throughput as the potential function. With the spatial uncoupling and the carefully designed potential function, the formulated game is shown to be an EPG. On the basis, the existence of NE solution of the considered game, which is locally optimal for the channel selection problems, is proved analytically.
3) To accomplish the self-learning during shared access, a reinforcement learning (RL) scheme is used, by designing a new decentralized Q-learning algorithm. Relying on the partial feedback information and the interaction with wireless environments, our scheme can achieve the NE points, by effectively excluding the frequent information exchanging among different links. Meanwhile, a rigorous proof of the convergence performance of this decentralized Q-learning algorithm is provided, which is shown to be more efficient than classical schemes.
4) The performances of our temporal-spatial reuse scheme are evaluated in mm-Wave UDN scenarios. First, in comparison with other schemes, the convergence performance of the new algorithm can be promoted effectively, in terms of both speed and stability, which leads also to a small latency. Second, by reusing spectrum via the spatial separation, the network throughput can be improved significantly. Meanwhile, the maximum accommodated devices in a local area are increased, which hence provides a great potential to 5G communications with dramatically increasing devices.

The rest of the article is organized as follows. In Section II, a new DSS model and its problem formulation are presented. In Section III, by designing the utility function, we formulate the DSS procedure as a non-cooperative game, and furthermore, investigate the properties of NE. In Section IV, we propose the decentralized Q-learning algorithm, and provide a rigorous proof of its convergence. Numerical simulations and performance evaluations are provided in Section V. The conclusions are made in Section VI.

II. System Model

A. System Model

We consider a UDN scenario with temporal-spatial reuse, which consists of K incumbent transmitter-receiver (TR) pairs and N shared TR pairs. There are M orthogonal channels available for use. For simplicity, we refer to each TR pair as one user. The licensed channels are possessed by incumebents/PUs and can be opportunistically used by SUs,

subject to no harmful interference. As shown in Fig. 1, multiple PU/SU links of UDN share the spectrum, which are randomly located in an area (e.g., an indoor office). In particular, the mm-Wave UDN is considered in which, (1) the strong path-loss reduces the risk of interferences [28], and (2) each user equipment (UE) is equipped with multiple antennas enabling spatial beam-forming [29], [30]. Denote the set of PU as $\mathcal{K}$, i.e., $\mathcal{K} = \{1, 2, ..., K\}$, the set of the SU as $\mathcal{N}$, i.e., $\mathcal{N} = \{1, 2, ..., N\}$, and the set of the licensed channels as $\mathcal{M}$, i.e., $\mathcal{M} = \{1, 2, ..., M\}$. In order to support the data transmission of PUs and reflect the spectrum resource competition among SUs, we assume $K \leqslant M \leqslant N$.

With the involvement of access point (AP), both spectrum occupancy status and locations of the PU will be available, for example, with the help of DS techniques. We consider that multiple competitive links are located on a 2-D grid, and the coordinate vector of the kth PU link is $\mathbf{t}_k^P = (x_k, y_k)$ (transmitter end), and $\mathbf{r}_k^P = (x_k, y_k)$ (receiver end), $k = 1, 2, ..., K$. The coordinate vector of the nth SU link is given by $\mathbf{t}_n^S = (x_n, y_n)$ (transmitter point), and $\mathbf{r}_n^S = (x_n, y_n)$ (receiver point), $n = 1, 2, ..., N$. Each SU transmitter will steer its beams based on a predefined beam codebook [31], [32], which will be aligned to a target receiver to enhance its SINR and sententiously avoid the exclusive regions PUs' receivers located. The width of main-lobes, or the beam resolution, is denoted by θ_n.

Given the above formulation, a state vector is used to characterize the nth SU:

$$\mathbf{s}_n = \left(\mathbf{t}_n^S, \mathbf{r}_n^S, \theta_n, \mathbf{D}, \mathbf{A}, \mathbf{G}\right). \tag{1}$$

Next, we elaborate on the later three parameters, which are used to describe co-interference relationships.

(1) **Euclidean distance matrix (EDM)**, i.e.,

$$\mathbf{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,K} & \cdots & d_{1,K+N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,K} & \cdots & d_{2,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{K,1} & d_{K,2} & \cdots & d_{K,K} & \cdots & d_{K,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{K+N,1} & d_{K+N,2} & \cdots & d_{K+N,K} & \cdots & d_{K+N,K+N} \end{bmatrix} \tag{2}$$

which specifies the distance between the transmitter and receiver (or both PU and SU), i.e., $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, $i, j \in \mathcal{K} \cup \mathcal{N}$. It is noted that this EDM will be only required in mathematically defining the link SINR, which in implementations needs not to be estimated.

(2) **Angle matrix (AM)**, i.e.,

$$\mathbf{A} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,K} & \cdots & \alpha_{1,K+N} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,K} & \cdots & \alpha_{2,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{K,1} & \alpha_{K,2} & \cdots & \alpha_{K,K} & \cdots & \alpha_{K,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{K+N,1} & \alpha_{K+N,2} & \cdots & \alpha_{K+N,K} & \cdots & \alpha_{K+N,K+N} \end{bmatrix} \tag{3}$$

which gives the direction of the line between transmitters and receivers. As mentioned, as the partial information, the direction of PU's receivers can be available at each SU transmitter. While the direction of other SUs, similar to the above EDM, will not be explicitly required, as the SINR will be estimated at each receiver and reported to the aligned transmitter.

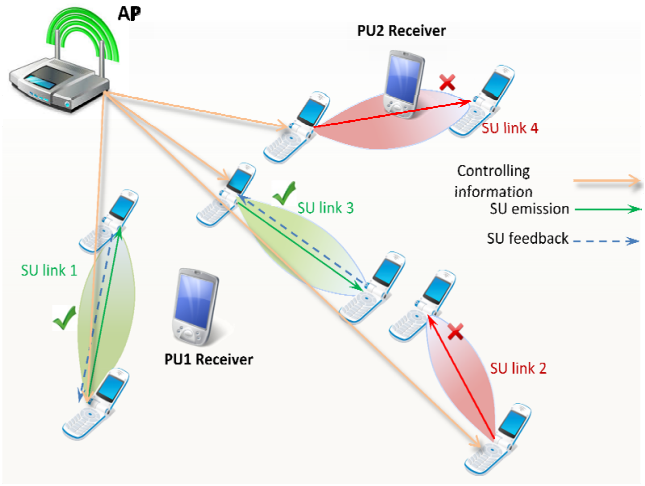

Fig. 1: The system model of mm-Wave UDN, here the few links are plotted for an illustrative purpose.

(3) **Beam Gains matrix (BGM)**, i.e.,

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,K} & \cdots & g_{1,K+N} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,K} & \cdots & g_{2,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{K,1} & g_{K,2} & \cdots & g_{K,K} & \cdots & g_{K,K+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{K+N,1} & g_{K+N,2} & \cdots & g_{K+N,K} & \cdots & g_{K+N,K+N} \end{bmatrix} \tag{4}$$

which gives the beam gains of different spatial angles with regards to its aligned direction (steering to its receiver). As specified in mm-Wave communications, the beam gain can be modeled as a circularly symmetric Gaussian function [33], with its maximum gain located at the target direction α_{ii}, i.e., $g_{i,j} = \exp\left\{-\frac{(\alpha_{ij} - \alpha_{ii})^2}{(\theta_i/30)^2 \times 50}\right\}$, $i, j \in \mathcal{K} \cup \mathcal{N}$.

Given the coordination signaling and the unaffordable latency, a full information exchanging among shared users will be infeasible in UDN. As mentioned, the local feedback scheme is thereby used, as in [34], which will facilitate the decision making of SUs by reducing the overhead signaling. To be specific, each SU receiver will measure its SINR and report it to the aligned SU transmitter, so that the transmitter can be aware of channel qualities. On the basis, SU will optimize its access strategy, and make reasonable adaption to channel selections in the next time slot.

A schematic structure is given by Fig. 2. Assume each SU is able to sense only one channel during a sensing period, and then select one channel for transmission in a slot [23]. At the beginning of each slot, each SU will choose a channel to sense, according to its selection strategy. If the selected channel is sensed by more than one SU, the mutual contention may occur. The channel access is successful in the case that the SINR is

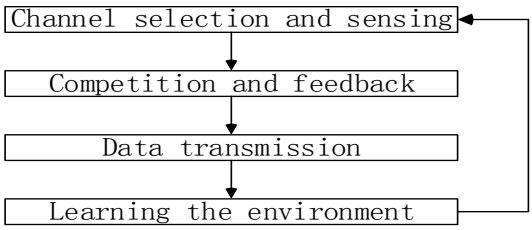

Fig. 2: Transmission structure of the SUs

not less than the predefined threshold, and otherwise, the SU has to keep silence in this slot and wait for the next slot.

Owing to the spatial separation, one channel can be now accessed by more than one SU, rendering the access process more complicated. To cope with it, we divide the competition procedure into two stages. In the first stage, SUs with the SINR above a predefined threshold are considered to be successful in the contention. As expected, there may be several successful SUs at this stage. The remaining SUs, with their SINR below this predefined threshold, will continue competing in the 2nd stage. In the following competition, only one SU can succeed. Then, successful SUs will transmit in the remaining slot. At the end of each slot, SUs will receive their rewards, and further learn wireless environments and update their access strategies.

After finite iterations, a whole network achieves its convergent state. For example, as in Fig. 1, the SU link 1 and link 3 are allowed to share spectrum with PU, as they cause no interference to PUs. In contrast, the SU link 2 and link 4 are denied to emitting. This is because, the SU link 2 will cause co-channel interference and thereby deteriorate network capacity, while the SU link 4 will cause intolerable interference to incumbent receiver.

B. Propagation Model

For mm-Wave communications, the propagation path-loss will be related to both distance and frequency [35], [36]. For simplicity, here we focus on the distance-dependent path-loss model (i.e., neglecting the effect from frequency), which is expressed as:

$$PL(d)[dB] = PL(d_0)[dB] + 10\kappa \log_{10}(d/d_0) \times \Big(1 - H\log_2(B_r)\Big) + X_\iota, \quad (5)$$

where $d_0 = 1m$ is the reference distance, $\kappa \in [2.2, 2.5]$ (e.g. LoS scenarios of 60GHz band) is the path loss exponent (PLE) for the strongest beam, $H = 0.06$ is the weighting factor, B_r is the number of unique pointing beams combined, and $X_\iota \in [8.2, 10.6]$ is the typical log-normal random shadowing variable. $PL(d_0)[dB]$ represents the path loss at a reference distance d_0, i.e.,

$$PL(d_0)[dB] = 20\log_{10}\frac{4\pi d_0}{\lambda} = 32.4 + 20\log_{10}(f_{GHz}). \quad (6)$$

Let $s_t(n)$ denote the transmitted signal of SU n in the mth channel. The received signal of SU n follows:

$$s_r(n) = \sqrt{P_{K+n}}s_t(n) + \sum\nolimits_{i \in \mathcal{K}\cup\mathcal{N}\setminus n} \sqrt{P_i}s_t(i) + w_m, \quad (7)$$

where P_i is the received power subject to channel propagations and beam gains, $i \in \mathcal{K} \cup \mathcal{N}$, and w_m is the additive white Gaussian noise (AWGN) with the zero mean and a variance of σ_m^2, i.e. $w_m \sim \mathcal{N}(0, \sigma_m^2)$ and $m \in \mathcal{M}$.

C. Problem Formulation

With the predefined angle matrix $\mathbf{A}$ and the beam gain matrix $\mathbf{G}$, we denote the spatial region interfered by the nth SU as $\mathcal{B}_n$, i.e.,

$$\mathcal{B}_n \triangleq \left\{\beta : \beta \in \left[\alpha_{K+n,K+n} - \frac{\theta_n}{2}, \alpha_{K+n,K+n} + \frac{\theta_n}{2}\right]\right\}, \quad \forall n \in \mathcal{N}. \quad (8)$$

Accordingly, denote these PUs suffering from the interference of nth SU with $\mathcal{I}_n$, i.e.,

$$\mathcal{I}_n \triangleq \Big\{i \in \mathcal{K} : \alpha_{K+n,i} \in \mathcal{B}_{K+n}\Big\}, = \Big\{i \in \mathcal{K} : g_{K+n,i} > g_0\Big\}, \quad (9)$$

where g_0 is the threshold of the beam gain and configured to $g_0 = 0.001$, which means the interference of user i to user j is trivial, if the beam gain $g_{i,j}$ is no greater than 0.001. Similarly, those SUs that will interfere to the nth SU is denoted as $\mathcal{J}_n$, i.e.,

$$\mathcal{J}_n \triangleq \Big\{j \in (\mathcal{N} \setminus n) : \alpha_{K+j,K+n} \in \mathcal{B}_{K+j}\Big\}, = \Big\{j \in (\mathcal{N} \setminus n) : g_{K+j,K+n} > g_0\Big\}. \quad (10)$$

In the considered UDN scenario, the interference accumulation from multiple SUs can be hardly controlled, and the null interference constraint becomes an alternative solution. Let A_k denote the channel occupied by PU k, hence the channel selection set for SU n is denoted by $\mathcal{A}_n$, i.e.,

$$\mathcal{A}_n \triangleq \Big\{m \in \mathcal{M} : m \neq A_k, k \in \mathcal{I}_n\Big\} \cup \varnothing, \quad (11)$$

and the channel selection of SU n is denoted by a_n, $a_n \in \mathcal{A}_n$. Further, the interference of SU n choosing the channel a_n, which is aroused by other users who belong to the set $\mathcal{J}_n$ and select the same channel a_n, is denoted by $I_{a_n}^n$, i.e.,

$$I_{a_n}^n = \sum\nolimits_{j \in \mathcal{J}_n} P_j \delta(a_n, a_j), \quad (12)$$

where $\delta(a_n, a_j)$ is the indicator function, i.e.,

$$\delta(a_n, a_j) = \begin{cases} 1, & a_n = a_j, \\ 0, & a_n \neq a_j. \end{cases}$$

Accordingly, the received SINR of the nth SU choosing the channel m can be denoted by:

$$\gamma_{n,m} = \frac{P_{K+n}}{I_m^n + \sigma_m^2}. \quad (13)$$

As we stated, the SINR value $\gamma_{n,m}$ of SU n determines the channel competition result. Let $b_n(\gamma_{n,m})$ indicate whether SU n successfully accesses the channel m or not after contention. Clearly, $b_n(\gamma_{n,m})$ is a Bernoulli random variable, i.e.,

$$b_n(\gamma_{n,m}) = \begin{cases} 1, & \gamma_{n,m} \geqslant \gamma_0, \\ 0, & \gamma_{n,m} < \gamma_0, \end{cases} \quad (14)$$

where γ_0 is the predefined SINR threshold.

It is assumed that all channels supply the same bandwidth W to each user, whilst different users may experience various channel qualities due to complex co-interferences. The achievable shared capacity is determined by the Shannon's formula, i.e., the capacity $r_{n,m}$ of the SU n achieved by accessing channel m is given by:

$$r_{n,m} = \begin{cases} W\log_2(1+\gamma_{n,m}), & b_n(\gamma_{n,m})=1, \\ 0, & b_n(\gamma_{n,m})=0. \end{cases} \quad (15)$$

Moving on, the expected reward, i.e., the effective capacity achieved by SU n ($n \in \mathcal{N}$) accessing channel a_n is

$$\begin{aligned} \bar{r}_{n,a_n} &= \mathbb{E}\Big(r_{n,a_n} b_n(\gamma_{n,a_n})\Big) \\ &= W\log_2(1+\gamma_{n,a_n})\mathbb{E}\Big(b_n(\gamma_{n,a_n})\Big). \end{aligned} \quad (16)$$

The network throughput, or the aggregate channel capacities obtained by all shared SUs, is then given by:

$$U(\mathbf{a}) = \sum_{n=1}^{N} \bar{r}_{n,a_n}, \quad (17)$$

where $\mathbf{a} = \{a_1, a_2, a_3, ..., a_N\}$ is a channel selection pattern for all SUs, with $a_n \in \mathcal{A}_n$.

When two or more SUs choose the same channel, then the mutual interference may occur even if the spatial beams are adopted. Therefore, in order to mitigate such co-channel interference among SUs, it is desirable to optimize the channel allocation for SUs, with the objective of maximizing the accumulated throughput. Thus, the channel access in UDN will be formulated as:

$$\mathcal{P} : \max U(\mathbf{a}). \quad (18)$$

It is seen that solving the above eq. (18) is challenging, as there is no centralized coordinator, and the complete or global information is unavailable (e.g. the positions of other SUs remain unknown). Thus, a distributed approach with a self-learning ability and low-complexity implementation will be of great importance.

III. Noncooperative Game Model for Dynamic Spectrum Sharing

Since there is no centralized coordinator, the channel selection has to be realized by each SUs independently premised on the uncompleted information. As a powerful tool for distributed decision making, where the individual decision will mutually influences each other, the game theory has been widely applied [17]–[19], which is suitable for traditional CR networks [37], [38]. One can refer to ref. [39] for various game-approach methods. The justification of applying a game-theoretic approach to the formulated problem are two-folds. First, SUs are both rational and selfish, which will make decision independently. Second, the objectives of multiple SUs may become conflicting while their decisions are achieved interactively.

In this section, an effective strategy for the game approach is designed in UDN scenarios, in which the global information (e.g. of other SUs) remains unknown. Note that, the main objective of this new scheme is to maximize the profits of SUs, relying on the concept of equilibrium.

A. Strategy Form Game

Some fundamental definitions are presented in the following. First, we establish the interference relationship as a graph.

*Definition 1 (**Interference Graph**):* We define $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ as an interference graph, where

- $\mathcal{N}$ is the set of players (SUs), $\mathcal{N} = \{1, 2, ..., N\}$;
- $\mathcal{E}$ is the set of edges, which is defined as $\mathcal{E} = \big\{(\mathbf{r}_i, \mathbf{t}_j) | i \in \mathcal{N}, j \in \mathcal{J}_i\big\}$. Note that, here we describe it from a perspective of receivers.

Given the interference graph, the game strategy can be then formulated as follows. For the notational convenience, the set of actions of all SUs, except the nth SU, is denoted as $\mathcal{A}_{-n}$, with its element a_{-n}.

*Definition 2 (**Strategy Form Game**):* A strategy form game is described as $\mathcal{F} = (\mathcal{G}, \mathcal{A}, \mathbf{u})$, in which

- $\mathcal{G}$ is interference graph, as we have defined previously;
- For each player $n \in \mathcal{N}$, $a_n \in \mathcal{A}_n$ is one action, and $\mathcal{A}_n$ is the set of feasible actions (channel selection) of the player n. Then, a pure strategy selection pattern is a n-tuple $\mathbf{a} = \{a_1, a_2, a_3, ..., a_N\}$, and the set of actions is $\mathcal{A} = \otimes\mathcal{A}_n$, where $\otimes$ is the Cartesian product;
- $\mathbf{u} = \{u_1, u_2, u_3, ..., u_N\}$ is the set of utility functions for the players, where $u_n(a_n, a_{-n})$ is the utility function of SU n, $a_n \in \mathcal{A}_n$, $a_{-n} \in \mathcal{A}_{-n}$.

To summarize, the proposed strategy is a kind of game, in which the utility of a player is not only dependent of the actions of itself, but also the actions of other players.

Then, we define the NE as follows, which accounts for the steady state of a non-cooperative game.

*Definition 3 (**Nash Equilibrium**):* An action pattern $\mathbf{a}^{NE} = \{a_1^{NE}, a_2^{NE}, a_3^{NE}, ..., a_N^{NE}\}$ is a pure strategy NE, if and only if no player can improve its utility by deviating unilaterally, i.e.,

$$\begin{aligned} u_n(a_n^{NE}, a_{-n}^{NE}) \geqslant & u_n(a_n, a_{-n}^{NE}), \\ & \forall n \in \mathcal{N},\ \forall a_n \in \mathcal{A}_n,\ a_n \neq a_n^{NE}. \end{aligned} \quad (19)$$

Finally, we present the definition of EPG which guarantees the existence of NE.

*Definition 4 (**Exact Potential Game**):* A game is an EPG if there exists an ordinal potential function $\Phi : \mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_n$ such that for all $n \in \mathcal{N}, a_n \in \mathcal{A}_n, a_n^* \in \mathcal{A}_n, a_{-n} \in \mathcal{A}_{-n}$, the following relation holds:

$$u_n(a_n, a_{-n}) - u_n(a_n^*, a_{-n}) = \Phi(a_n, a_{-n}) - \Phi(a_n^*, a_{-n}). \quad (20)$$

The change in a utility function caused by the unilateral action change of an arbitrary player is exactly the same with that in the potential function. Thus, EPG belongs to the potential games, which have been applied widely to wireless communications. Potential game exhibits several attractive properties and two most important of them are:

- Every potential game has at least one pure strategy NE;
- any global or local maxima of the potential function will constitute a pure strategy NE.

B. Utility Function

In existing works, the utility functions are commonly selected as a group of indicator functions, suggesting if a player competes successfully with its current action, then it will acquire a unit reward 1 and, otherwise, it will obtain a reward of 0. In the context of shared access in UDN, we alternatively formulate the utility function of player n as its attained channel capacity in the strategy-form game, i.e.,

$$u_n(a_n, a_{-n}) \triangleq \mathbb{E}\Big(r_{n,a_n}|a_{-n}\Big) = \bar{r}_{n,a_n}, \tag{21}$$

where r_{n,a_n} is the random reward received by player n when taking the action a_n.

Based on the above analysis and the rational and selfish nature of players, we formulate the problem of maximizing network throughput in eq. (18) as a distributed strategy-form game, which can be further expressed as:

$$\mathcal{F}: \max_{a_n \in \mathcal{A}_n} u_n(a_n, a_{-n}), \forall n \in \mathcal{N}, \tag{22}$$

where $\mathcal{A}_n$ is the action set (i.e., the available channel set) of player n specified by eq. (11).

In the following, we first prove the formulation optimization problem in eq. (22) is an EPG. Thus, the NE will correspond to its optimal solution. On this basis, we design a decentralized Q-learning algorithm to obtain the NE solution.

C. Analysis of Nash Equilibrium

We firstly investigate the properties of the above strategy-form game, with which the existence and convergence of NE can be demonstrated.

Theorem 1: The considered game $\mathcal{F}$ in eq. (22) is an EPG, which has at least one pure strategy NE. In addition, the optimal solution to the problem $\mathcal{P}$ in eq. (18), i.e., maximizing the throughput of UDN, constitutes a pure strategy NE of $\mathcal{F}$.

Before proceeding, we define the potential function as:

$$\Phi(a_n, a_{-n}) = \sum\nolimits_{m=1}^{M} \sum\nolimits_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big(\mathcal{C}_m(c)\Big), \tag{23}$$

where $|\mathcal{C}_m|$ denote the number of SUs who select channel m for sense and competition, i.e., $\mathcal{C}_m = \{n \in \mathcal{N} : a_n = m\}$ $(m \in \mathcal{M})$, and

$$\varphi_m\Big(\mathcal{C}_m(c)\Big) \triangleq W \log_2(1 + \gamma_{\mathcal{C}_m(c),m}) \mathbb{E}\Big(b_{\mathcal{C}_m(c)}(\gamma_{\mathcal{C}_m(c),m})\Big), \\ c = \Big\{1, 2, ..., |\mathcal{C}_m|\Big\}, m \in \mathcal{M}. \tag{24}$$

The proof procedure will be divided into two parts. First, we consider a simple yet fundamental instance, in which the SU n will not suffer from interference of other SUs. Second, we will generalize it to a common situation, where the SU n suffers from mutual interference.

Assume that the casual SU n changes unilaterally its selected channel from a_n to a_n^*, and the resulting variation in individual utility function caused by this one-sided adaption is written as:

$$\begin{aligned} &u_n(a_n, a_{-n}) - u_n(a_n^*, a_{-n}) = \bar{r}_{n,a_n} - \bar{r}_{n,a_n^*}, \\ &= W \log_2(1 + \gamma_{n,a_n}) \mathbb{E}\Big(b_n(\gamma_{n,a_n})\Big) \\ &\quad - W \log_2(1 + \gamma_{n,a_n^*}) \mathbb{E}\Big(b_n(\gamma_{n,a_n^*})\Big), \\ &= \varphi_{a_n}(n) - \varphi_{a_n^*}(n). \end{aligned} \tag{25}$$

For the simple situation, as the player n is free from interferences of other players, the situation is simple and intuitive, i.e., the change in the potential function caused by

$$\begin{aligned} &\Phi(a_n, a_{-n}) - \Phi(a_n^*, a_{-n}) \\ &= \sum\nolimits_{m=1}^{M} \sum\nolimits_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big[\mathcal{C}_m(c)|(a_n, a_{-n})\Big] - \sum\nolimits_{m=1}^{M} \sum\nolimits_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big[\mathcal{C}_m(c)|(a_n^*, a_{-n})\Big], \\ &= \left[\sum\nolimits_{m \in \mathcal{M} \setminus (a_n \cup a_n^*)} \sum_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big(\mathcal{C}_m(c)\Big) + \sum_{c=1}^{|\mathcal{C}_{a_n}|} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big) + \sum_{c=1}^{|\mathcal{C}_{a_n^*}|} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big)\right] \\ &\quad - \left[\sum\nolimits_{m \in \mathcal{M} \setminus (a_n \cup a_n^*)} \sum_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big(\mathcal{C}_m(c)\Big) + \sum_{c=1}^{|\mathcal{C}_{a_n}|-1} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big) + \sum_{c=1}^{|\mathcal{C}_{a_n^*}|+1} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big)\right], \\ &= \left[\sum_{c=1}^{|\mathcal{C}_{a_n}|} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big) - \sum_{c=1}^{|\mathcal{C}_{a_n}|-1} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big)\right] + \left[\sum_{c=1}^{|\mathcal{C}_{a_n^*}|} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big) - \sum_{c=1}^{|\mathcal{C}_{a_n^*}|+1} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big)\right], \\ &\overset{(a)}{=\!=} \left[\left(\sum_{i \in \mathcal{C}_{a_n} \setminus n} \varphi_{a_n}(i) + \varphi_{a_n}(n)\right) - \sum_{c=1}^{|\mathcal{C}_{a_n}|-1} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big)\right] + \left[\sum_{c=1}^{|\mathcal{C}_{a_n^*}|} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big) - \left(\sum_{i \in \mathcal{C}_{a_n^*}} \varphi_{a_n^*}(i) + \varphi_{a_n^*}(n)\right)\right], \\ &\overset{(b)}{=\!=} \varphi_{a_n}(n) - \varphi_{a_n^*}(n). \end{aligned} \tag{27}$$

this unilateral change is given by:

$$\begin{aligned}
&\Phi(a_n, a_{-n}) - \Phi(a_n^*, a_{-n}) \\
&= \sum\nolimits_{m=1}^{M} \sum\nolimits_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big[\mathcal{C}_m(c)|(a_n, a_{-n})\Big] \\
&\quad - \sum\nolimits_{m=1}^{M} \sum\nolimits_{c=1}^{|\mathcal{C}_m|} \varphi_m\Big[\mathcal{C}_m(c)|(a_n^*, a_{-n})\Big], \\
&= \sum_{n=1}^{N} \Big[\bar{r}_{n,a_n}|(a_n, a_{-n})\Big] - \sum_{n=1}^{N} \Big[\bar{r}_{n,a_n}|(a_n^*, a_{-n})\Big], \\
&= \left[\sum\nolimits_{i\in\mathcal{N}\setminus n} \bar{r}_{i,a_i} + \bar{r}_{n,a_n}\right] - \left[\sum\nolimits_{i\in\mathcal{N}\setminus n} \bar{r}_{i,a_i} + \bar{r}_{n,a_n^*}\right], \\
&= \bar{r}_{n,a_n} - \bar{r}_{n,a_n^*} = \varphi_{a_n}(n) - \varphi_{a_n^*}(n). \qquad (26)
\end{aligned}$$

Then, we focus on the common case. The resulting change in potential function caused by this unilateral change is given by eq. (27). From the sharing point of view, the change of the player n's channel selections only influences the users within channels a_n and a_n^*. Based on this consideration, we divide the channel set into three types, i.e., $m \in \mathcal{M} \setminus (a_n \cup a_n^*)$, $m = a_n$ and $m = a_n^*$. On the other hand, due to the spatial separation, multiple users accessing the same channel successfully means there is no interference or the interference is insignificant among these SUs., which indicates the following equations hold.

$$\sum_{i\in\mathcal{C}_{a_n}\setminus n} \varphi_{a_n}(i) = \sum_{c=1}^{|\mathcal{C}_{a_n}|-1} \varphi_{a_n}\Big(\mathcal{C}_{a_n}(c)\Big),$$

$$\sum_{c=1}^{|\mathcal{C}_{a_n^*}|} \varphi_{a_n^*}\Big(\mathcal{C}_{a_n^*}(c)\Big) = \sum_{i\in\mathcal{C}_{a_n^*}} \varphi_{a_n^*}(i).$$

Therefore the eq. (27) from (a) to (b) can be guaranteed. Note that, the eq. (26) is a special case of eq. (27).

From eqs. (25)-(27), we immediately reach the following equation:

$$\Phi(a_n, a_{-n}) - \Phi(a_n^*, a_{-n}) = u_n(a_n, a_{-n}) - u_n(a_n^*, a_{-n}). \quad (28)$$

From eq. (28), we note that the change in individual utility function, caused by an arbitrary player's unilateral declination, is identical with the change in the potential function, which confines to the definition of EPG, as in eq. (20). Therefore, the formulated shared game $\mathcal{F}$ is an EPG, employing the aggregated network throughput as a potential function, which hence has at least one pure strategy NE. Thus, according to the relationship between the potential function eq. (23) and the objective function eq. (18), *Theorem 1* can be proved.

IV. Decentralized Learning

In order to achieve the optimal NE points of the above strategy-form game, a new distributed adaption algorithm is designed, premised on the RL concept [40]. Due to the consideration of complexity, we cannot make the assumption on channel selection probabilities of SUs. We solve this problem by way of multiple self-learning processes, which aim at independently adapting each SU's access strategy. It is noteworthy that, rather than the global and complete information of wireless environments, only partial information can be available, i.e., the individual history of each user's decisions and rewards as well as PU's exclusive region. Thus, we suggest a decentralized Q-learning algorithm, which can maximize the aggregated channel capacity of UDN. In this regards, each SU will learn environments from the individual action-reward experience and adjust their selection strategies that will finally converge to a NE point.

A. Decentralized Q-learning

To facilitate the elaboration on the scheme, we expand the game $\mathcal{F}$ of shared access into a mixed strategy form game. Let $\mathbf{P}$ denote the mixed strategy probability of the formulated game, i.e.,

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,M} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,M} \\ \vdots & \vdots & p_{n,m} & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,M} \end{bmatrix}, \qquad (29)$$

where $p_{n,m}$ denotes the probability of nth SU selecting channel m. Let $\mathbf{p}_n = (p_{n,1}, p_{n,2}, ..., p_{n,M})$ denote the probability vector of nth SU selecting actions, and intuitively, we have $\sum_{m=1}^{M} p_{n,m} = 1$.

We define the Q-function as the expected reward of each SU under different actions, That is, for SU n, the Q-value of choosing channel a_n in k slot is:

$$Q_{n,a_n}(k) = r_{n,m}|\big(m = a_n(k)\big), \qquad (30)$$

where $r_{n,m}$ is the reward of SU n, which is dependent on actions taken both by player n and other players, as well as the PUs' status.

In the Q-learning procedure, Q-values of time k will be updated on that of previous time (k-1):

$$Q_{n,m}(k+1) =$$

$$\begin{cases} Q_{n,m}(k), & \text{if } m \neq a_n(k), \quad (31a) \\ \Big[1 - \xi_{n,m}(k)\Big] Q_{n,m}(k) + \xi_{n,m}(k) \cdot r_{n,m}(k), & \text{else}, \quad (31b) \end{cases}$$

where $\xi_{n,m}(k) = \frac{1}{k+2}$ is the step factor, which are expected to meeting:

- $0 \leqslant \xi_{n,m}(k) < 1$, that is $0 \leqslant \dfrac{1}{k+2} < 1$, i.e., $k \geqslant 0$.
- $\sum\limits_{k=0}^{\infty} \xi_{n,m}(k) = \infty, \forall n \in \mathcal{N}, m \in \mathcal{M}$.

In contrast to the fictitious play that will be deterministic, the action in Q-learning algorithm is randomly taken, such that all actions will be tested. Here, we are interested in the Boltzmann distribution for random explorations, and the expression of the element $p_{n,m}$ of the mixed strategy probability $\mathbf{P}$ is given by:

$$p_{n,m} = \frac{e^{Q_{n,m}(k)/\rho}}{\sum\limits_{i=1}^{M} e^{Q_{n,i}(k)/\rho}}, \qquad (32)$$

where ρ is referred to a virtual temperature, which controls the frequency of exploration. Obviously, the smaller ρ is (the colder), the more focused the actions are. When $\rho \to 0$, each user tends only to choose the channel with the largest Q-value.

Algorithm 1 Decentralized Q-learning algorithm for player n

Step 1: Set $k = 0$ and the initial channel selection probability vector $p_{n,m} = 1/M$, $Q_{n,m}(k) = 0$, $\forall n \in \mathcal{N}$, $\forall m \in \mathcal{M}$.
Step 2: At the beginning of the k slot, each SU n chooses a channel $a_n(k)$ according to its current channel selection probability vector $\mathbf{p}_n(k)$. In each slot, the SUs perform channel sensing and contention which contains two stage. In the first stage, the SUs those SINR above the threshold are perceived as successes and complete their contention, and the remaining SUs continue competing with those SINR below the threshold in the second stage.
Step 3: Calculate the reward of each SU according to eq. (15).
Step 4: All the SUs update their Q-value according to the rules that described as eqs. (31).
Step 5: Then update their channel selection probability $p_{n,m}(k)$ according to the expression that given by eq. (32).
Step 6: $\forall n \in \mathcal{N}$, if there exists a component $p_{n,m}$ of $\mathbf{p}_n(k)$ which is approaching one, e.g., larger than 0.99, stop; Otherwise, go to step 2.

The decentralized Q-learning algorithm, suggested for two-dimension sharing access in UDN, is summarized in **Algorithm 1**. Note that, two points are worth highlighting to our new sharing scheme. First, the complete information about actions taken by other players can be excluded at each shared link, and therefore, the channel selection probability vector $\mathbf{p}_n$ will be updated on its own reward from the competitive environments. Specifically, if a channel is selected and the reported SINR surpasses the predefined threshold, i.e. $\gamma_{n,m} \geqslant \gamma_0$, the probability of selecting this channel in the next slot will increase. Reversely, if the SINR is below the threshold, i.e. $\gamma_{n,m} < \gamma_0$, the probability of selecting this channel in the next slot remains unchanged. Second, the reward of SU of each iteration is equivalent to the achieved channel capacity (rather than binary indicator values 1 and 0), which conforms more to the reality and can facilitate the convergence to some extents.

For each player n ($n \in \mathcal{N}$), once a component $p_{n,m}$ in the channel selection probability vector $\mathbf{p}_n$ is sufficiently larger, for example surpassing 0.99, then the self-adaption algorithm will be terminated, and a group of actions be derived.

B. Convergence Analysis

It is understood that the faster the convergence of iterative algorithms, the higher the efficiency of both time and spectrum. Thus, the convergence speed of the learning scheme is of promise to shared access, especially in dynamic environments. Then, we will discuss the convergence of the decentralized Q-learning algorithm in UDN scenarios with the temporal-spatial reuse. First, the existence of convergence state of the algorithm is proved, by resorting to the concept of stochastic approximations [41]. Second, we will study the convergence performance in terms of speed and stability.

1) Existence of Convergence: The Q-values for different interfered SUs are mutually coupled, thus all Q-values will change once a single Q-value is changed. We will refer to Q-values meeting the following condition as a *stationary* point.

$$Q_{n,a_n}(k) = r_{n,a_n}(k) \times \Pr\left[\prod_{j\in\mathcal{J}_n}\Big(a_j(k) \neq a_n(k)\Big)\right], \\ = r_{n,a_n}(k) \times \prod_{j\in\mathcal{J}_n}\left(1 - \frac{e^{Q_{j,a_n}(k)/\rho}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\rho}}\right). \quad (33)$$

Note that, the stationarity point holds only in a statistical sense, as the Q-values can undulate around because of the randomness of channel selections. It is shown that, with $\rho \to 0$, the stationary point will *converge* to a NE point. Unfortunately, we are still not sure about the existence of such a stationary point. Hereinafter, we provide a theorem guaranteeing the existence of stationary points and quote three lemmas to prove it.

Theorem 2: The Q-learning converges to a stationary point with the probability 1.

To prove this theorem, we recommend the following lemmas.

Lemma 1: For an adequately small ρ, there exists at least one stationary point satisfying eq. (33).

For analytical convenience, we define:

$$\mathbf{q} \triangleq \Big(Q_{11}, ..., Q_{1M}, Q_{21}, ..., Q_{2M}, ..., Q_{N1}, ..., Q_{NM}\Big)^T, \quad (34)$$

$$\mathbf{r} \triangleq \Big(r_{11}, ..., r_{1M}, r_{21}, ..., r_{2M}, ..., r_{N1}, ..., r_{NM}\Big)^T, \quad (35)$$

and

$$h(\mathbf{q}) = \prod_{j\in\mathcal{J}_n}\left(1 - \frac{e^{Q_{j,a_n}(k)/\rho}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\rho}}\right), \quad (36)$$

thus, the above eq. (33) can be rewritten as:

$$f(\mathbf{q}) = \mathbf{q} - \mathbf{r}h(\mathbf{q}) = 0. \quad (37)$$

Then, the updating rule in eq. (31b) is equivalent to solving the equation eq. (33) using the Robbins-Monro algorithm [42], i.e.,

$$\mathbf{q}(k+1) = \Big[1 - \xi(k)\Big] \times \mathbf{q}(k) + \xi(k)\mathbf{r}(k), \\ = \mathbf{q}(k) + \xi(k) \times \Big[\mathbf{r}(k) - \mathbf{q}(k)\Big], \\ = \mathbf{q}(k) + \xi(k) \times \Big[\bar{\mathbf{r}}(k) - \mathbf{q}(k) + \mathbf{r}(k) - \bar{\mathbf{r}}(k)\Big], \\ = \mathbf{q}(k) + \xi(k) \times \Big[\bar{\mathbf{r}}(k) - \mathbf{q}(k) + \eta\mathbf{m}(k)\Big], \quad (38)$$

where $\bar{\mathbf{r}}(k) = \mathbb{E}[\mathbf{r}(k)]$, and $\eta\mathbf{m}(k) = \mathbf{r}(k) - \bar{\mathbf{r}}(k)$. Obviously, we have $\mathbb{E}\{\eta\mathbf{m}(k)\} = 0$. Therefore, the term $\eta\mathbf{m}(k)$ is a Martingale difference.

After checking eq. (31) and eq. (38), it is seen that the updating processing of Q-values is the exact stochastic approximation of the solution to eq. (38). It is well known that the convergence of such a procedure will be related with an ordinary differential equation (ODE). As $\eta\mathbf{m}(k)$ in eq. (38) is a Martingale difference, it is easy to obtain the following lemma:

Lemma 2: With probability 1, the sequence $\mathbf{q}$ will converge to some limit set of the ODE:

$$\mathbf{q} = f(\mathbf{q}). \tag{39}$$

What remains to do, equivalently, is to analyze the convergence property of the ODE in eq. (39). We then obtain the following lemma by applying the Lyapunov function [43], [44].

Lemma 3: The solution of ODE eq. (39) converges to the stationary point determined by eq. (37).

Finally, combining Lemmas 1, 2, and 3, *Theorem 2* can be proved.

2) Performance of Convergence : After the proof of the existence of stationary point, we now investigate the mixed strategy probability $\mathbf{P}$ to analyze the convergence speed and stability of our decentralized Q-learning algorithm in UDN.

Intuitively, the convergence speed depends on two factors. First, from a system level, the different mechanisms of shared access determine the number of accessible SUs in each time slot. In general, the more accommodated SUs in one slot, the faster convergence of the system can achieve. Second, at the user level, as in eq. (32) the changing rate of Q-value in each iteration, which is influenced by the reward function $r_{n,m}$, has also effects on the number of required iterations. In this regards, a large gradient decent will lead to a fast convergence rate.

In view of the above considerations, we present a rigorous proof of the convergence speed as follows. First, we make the following proposition of the channel stable selection state.

*Proposition 1 (**Stable Selection State**):*

If there is a component $p_{n,a_n}(k) \in \mathbf{p}_n(k)$ sufficiently approaching 1 while other components $p_{n,m}(k), m \in \mathcal{M}, m \neq a_n$ sufficiently small (i.e., approaching 0), we can say that the SU's channel selection will remain invariant.

Second, as an innovative concept introduced by this work to analyze the convergence, we introduce the domain of attraction.

*Definition 5 (**Domain of Attraction**):* We define the channel selection probability vector $\mathbf{p}_n$ of the SU n entering the domain of attraction, if the probability of choosing the channel m in the current iteration k is not less than the probability of choosing the same channel in the previous iteration $(k-1)$, i.e.,

$$p_{n,m}(k) \geqslant p_{n,m}(k-1), m = a_n(k). \tag{40}$$

The channel selection probability vector $\mathbf{p}_n$ of SU n, after entering the domain of attraction, will be quickly increased. And after some finite iterations, the probability p_{n,a_n} would approximate 1, and at the same time, the other probability $p_{n,m}(m \neq a_n)$ may approach 0. Once the probability p_{n,a_n} surpass the threshold (e.g. 0.99), then we consider that the convergence condition is satisfied. During this region, the stability and convergency speed can be promoted jointly via a *positive reinforcement* mechanism. That is to say, if the channel selection is preferable to the more stable action, the convergence can be speeded up and the required iteration will be reduced. In order words, the higher stability comes also with the faster convergence.

Then, we introduce the sufficient condition about the channel selection probability vector $\mathbf{p}_n$ in a domain of attraction.

Theorem 3: The probability vector $\boldsymbol{p}_n$ satisfies the domain of attraction, if $\boldsymbol{p}_n$ is a channel stable selection state meanwhile the SU has a positive reward with this channel selection probability vector $\boldsymbol{p}_n$ in k time slot, i.e., $r_{n,a_n}(k) > 0, a_n = \arg\max_{a_n \in \mathcal{A}_n} \boldsymbol{p}_n$.

We investigate the relationships of channel selection probability of two adjacent slots, i.e., $p_{n,a_n}(k-1)$ and $p_{n,a_n}(k)$, as

$$\begin{aligned}
\frac{p_{n,a_n}(k)}{p_{n,a_n}(k-1)} &= \frac{e^{Q_{n,a_n}(k)/\rho} \Big/ \sum_{m=1}^{M} e^{Q_{n,m}(k)/\rho}}{e^{Q_{n,a_n}(k-1)/\rho} \Big/ \sum_{m=1}^{M} e^{Q_{n,m}(k-1)/\rho}} = \frac{e^{Q_{n,a_n}(k)/\rho}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\rho}} \times \frac{\sum_{m=1}^{M} e^{Q_{n,m}(k-1)/\rho}}{e^{Q_{n,a_n}(k-1)/\rho}}, \\
&= \frac{e^{Q_{n,a_n}(k)/\rho} \times \left[e^{Q_{n,a_n}(k-1)/\rho} + \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k-1)/\rho} \right]}{\left[e^{Q_{n,a_n}(k)/\rho} + \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k)/\rho} \right] \times e^{Q_{n,a_n}(k-1)/\rho}}, \\
&= \frac{e^{Q_{n,a_n}(k)/\rho} \times e^{Q_{n,a_n}(k-1)/\rho} + e^{Q_{n,a_n}(k)/\rho} \times \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k-1)/\rho}}{e^{Q_{n,a_n}(k)/\rho} \times e^{Q_{n,a_n}(k-1)/\rho} + e^{Q_{n,a_n}(k-1)/\rho} \times \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k)/\rho}}.
\end{aligned} \tag{41}$$

$$\frac{e^{Q_{n,a_n}(k)/\rho} \times e^{Q_{n,a_n}(k-1)/\rho} + e^{Q_{n,a_n}(k)/\rho} \times \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k-1)/\rho}}{e^{Q_{n,a_n}(k)/\rho} \times e^{Q_{n,a_n}(k-1)/\rho} + e^{Q_{n,a_n}(k-1)/\rho} \times \sum_{m \in \mathcal{M} \setminus a_n} e^{Q_{n,m}(k)/\rho}} \geqslant 1. \tag{42}$$

shown in eq. (41). For the sake of achieving eq. (40), Eq. (41) needs to meet the condition given by eq. (42). That is:

$$e^{Q_{n,a_n}(k)/\rho} \times \sum\nolimits_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k-1)/\rho} \geqslant e^{Q_{n,a_n}(k-1)/\rho} \times \sum\nolimits_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k)/\rho}. \quad (43)$$

Furthermore, the situations indicating by eq. (43) can be categorized into two cases:

1) *Case I*: The SU n chooses the same channel at two adjacent slots, i.e. $a_n(k) = a_n(k-1)$, and obtains a positive reward, i.e.,

$$e^{Q_{n,a_n}(k)/\rho} > e^{Q_{n,a_n}(k-1)/\rho}, \quad (44a)$$

$$\sum_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k-1)/\rho} = \sum_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k)/\rho}. \quad (44b)$$

2) *Case II*: The SU n chooses channel $m, m \in \mathcal{M}, m \neq a_n(k-1)$ in the current iteration k, and obtains a zero reward, i.e.,

$$e^{Q_{n,a_n}(k)/\rho} = e^{Q_{n,a_n}(k-1)/\rho}, \quad (45a)$$

$$\sum_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k-1)/\rho} > \sum_{m\in\mathcal{M}\setminus a_n} e^{Q_{n,m}(k)/\rho}. \quad (45b)$$

After the solutions have entered into the domain of attraction, which is monotonically increased, there will always exist another better solution in the domain of attraction at the kth iteration, compared with that of the $(k-1)$th iteration. In this situation, owing to the rational and selfish nature of players, they will take the action that makes the most advantage to their own and maximizes their rewards. Consequently, the *Case II* tends to be unstable, which does not fulfill the circumscription of the domain of attraction. Stated thus, *Theorem 3* can be proved.

Premised on *Theorem 3*, we draw the conclusion that, for the UDN scenarios with temporal-spatial reuse, the channel selection probability $\mathbf{P}$ meets Proposition 1 and enters into a domain of attraction after few iterations, due to the spatial separation and the reduced mutual influences. By conducting the parallel adaption with little coupling in multiple links, the iterations to archive NE will be significantly reduced. In contrast, in order to satisfy Proposition 1, other existing works, with strong coupling among different links, need a large number of iterations and probably obtain instable strategies. As demonstrated by subsequent simulations, the new scheme will be more attractive, in terms of convergence, to the emerging UDN involving many devices and requiring low latency.

V. Numerical Simulations and Discussions

In the section, numerical results are provided to evaluate the performance of the new scheme in the context of mm-Wave UDN. In the simulations, all SUs are randomly located in a local area. Without loss of generality, we assume $|\mathcal{M}| = 5$ licensed channels are available. The main-lobe width θ_n and the predefined SINR threshold γ_0 can be adjusted, depending on specific applications.

In the following, we firstly illustrate the convergence performance of our new scheme (i.e., the fine-grained two-dimensional reuse) and other existing spectrum access schemes. Then, we will evaluate the system performances (i.e., the network throughput and the maximum accommodated SUs number) of our new scheme in a context of UDN.

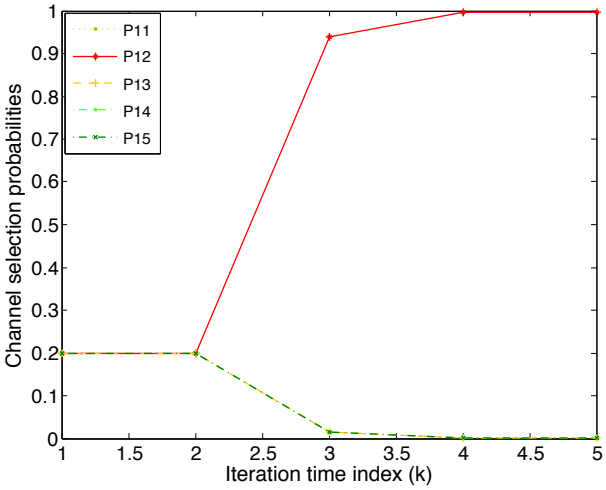

Fig. 3: The channel selection probabilities evolution of one SU with our proposed access scheme.

A. Convergence Performance

In the first simulation, the influence of different numbers of SU links on the UDN are evaluated. Here, the main-lobe beam width θ_n of user n and the predefined SINR threshold γ_0 are configured respectively to $\theta_n = \theta = 30°$, and $\gamma_0 = 10$dB. A counterpart scheme is simulated [23], where the mutual interference between SUs is intolerable and each channel may at most accommodate one SU.

First, the evolution curves of channel selection probabilities with our proposed access scheme is plotted in Fig. 3. From the results, it is seen that, after about 4 iterations, the domain of attraction will be attained, i.e. one of the channel selection probability $p_{1,2} \in \mathbf{p}_1$ has exceeded 0.99, whilst the other components in $\mathbf{p}_1$ have already approached 0. So, the convergence can be achieved after 4 iterations. Thus, owing to the spatial separation and effective learning mechanisms, our new scheme requires only little iterations to enter domain of attraction and thereby attains its convergence in UDN applications.

The cumulative distribution function (CDF) of the required iterations for convergence is shown in Fig. 4. In numerical derivations, 20 different network topologies are randomly generated, and for each network topology 500 independent trials are implemented. Together with the previous analysis, we can conclude that the proposed **Algorithm 1** shows the favorable converge performance in UDN applications. From Fig.4, given both 4 configurations (e.g. $N = 5$ for non-dense network and $N = 20$ for dense network), the iterations required for convergence of the new scheme will be dramatically less than that of other schemes, as also indicated our previous theoretical analysis, e.g. *Theorem 3*. More importantly, as the number of SU links increases from $N = 5$ to $N = 20$, the expected iterations is only increased from 4 to 28 in our new scheme, indicating a good scalability if applied to UDN. However, the

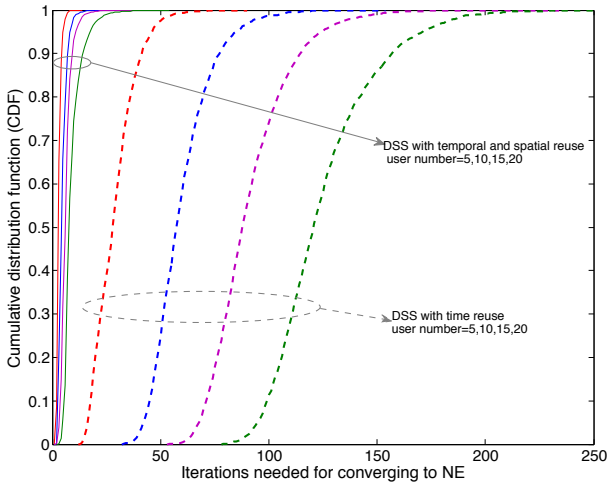

Fig. 4: The convergence speed comparison between the DSS with spatial reuse and the conventional DSS

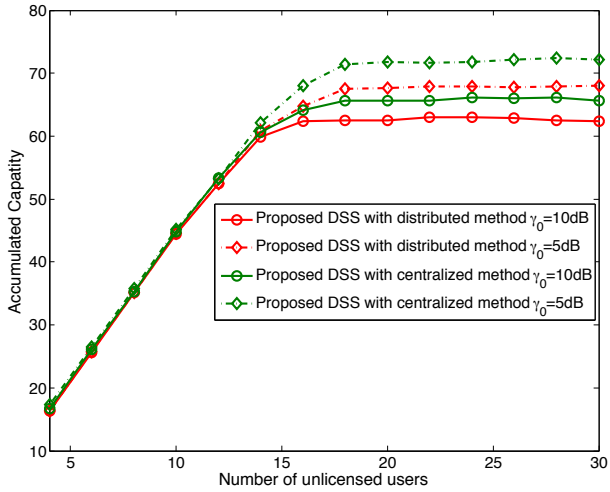

Fig. 6: Throughput comparison between the centralized and distributed algorithms

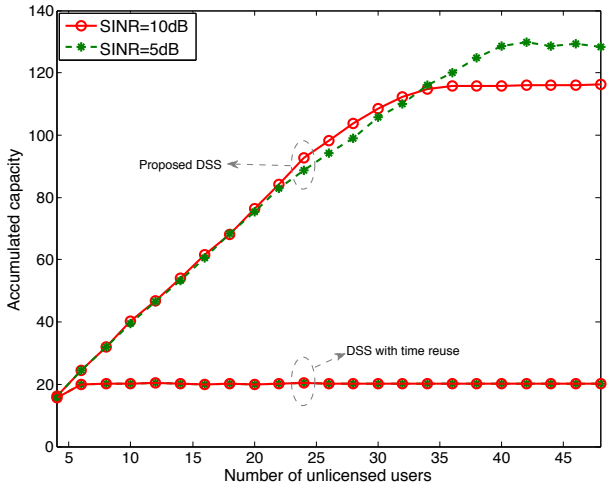

Fig. 5: Throughput comparison with the same main-lobe angle of transmitters while different threshold of SINR

required iterations of conventional schemes will be increased significantly from 40 to 200. Thus, our new scheme would incur the significantly reduced complexity in implementations. As far as the short-term spectrum opportunity and the access latency are concerned, the new access scheme with fine-grained reuse will be more attractive to UDN applications.

B. System Performance

In the following, we will evaluate the system throughput of the proposed two dimensional spectrum reuse scheme. To this end, the realistic influences on performance from three key parameters are comprehensively considered: (1) the main-lobe width θ_n, (2) the predefined SINR threshold γ_0, and (3) with the PUs emission status. Similarity, 20 randomly deployed network topologies are used with 500 independent trials in each topology.

1) The SINR Threshold: We firstly evaluate the effects of SINR threshold γ_0 on the network throughput. In order to illustrate the property of our two dimensional sharing mechanism with the distributed algorithm thoroughly, we divide the simulations into two parts, in which the comparison of our scheme with the single dimensional DSS is shown firstly, and then we present the performance of the centralized algorithm as a benchmark of our considered distributed algorithm.

For the first part, the main-lobe width $\theta_n = \theta = 30°$, and the number of total SU links varies from N=4 to N=50 for both temporal reuse method and two dimensional reuse scheme. Two typical configurations on the predefined SINR threshold are considered, i.e. $\gamma_0 = 5$dB and $\gamma_0 = 10$dB.

From Fig. 5, with the fine-grained sharing and the temporal-spatial reuse, it is found that the throughput of our new scheme will dramatically outperform the conventional schemes, under various predefined SINR thresholds γ_0. Meanwhile, as the number of SUs increasing in UDN, the performance gap between the new access scheme and conventional methods will also be increased gradually. This is easy to follow. In a conventional scheme, one channel admits only one SU in each slot and, otherwise, the collision among shared links will occur. In the new scheme, owing to the spatial separation and the lower coupling among shared links, one channel will be occupied harmoniously by more than one SU in each slot. Therefore, the network throughput will further grow, as the SUs number increases. For conventional access methods, however, the saturation will occur after the number of SUs surpasses the available vacant channels, by seriously restricting the UDN network throughput.

Besides, it is observed that the network throughput of a low SINR threshold ($\gamma_0 = 5$dB) may surpass that of a high one ($\gamma_0 = 10$dB). The main reason is that the number of accommodated SUs will be relatively different in the two cases. In general, the smaller of the predefined SINR threshold γ_0 is, the more SUs will be held. Note that, here will be a compromise between the network throughput and the quality of experience (QoE) in shared links. I.e. the lower SINR threshold indicates the lower transmission quality.

For the second part, the main-lobe width $\theta_n = \theta = 60°$, and the number of total SU links varies from N=4 to N=30

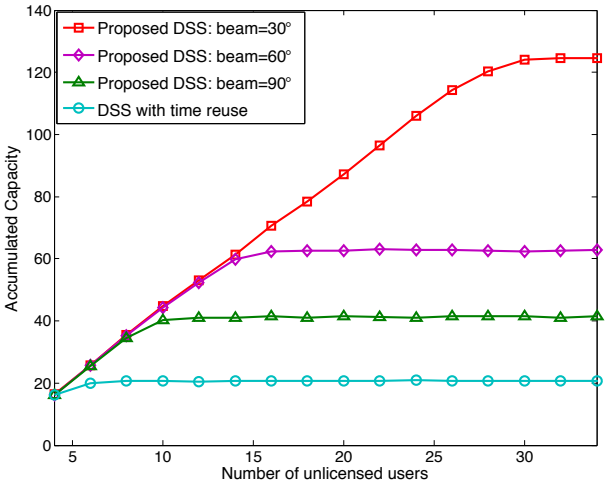

Fig. 7: Throughput comparison with the same threshold of SINR while different main-lobe angle of transmitters

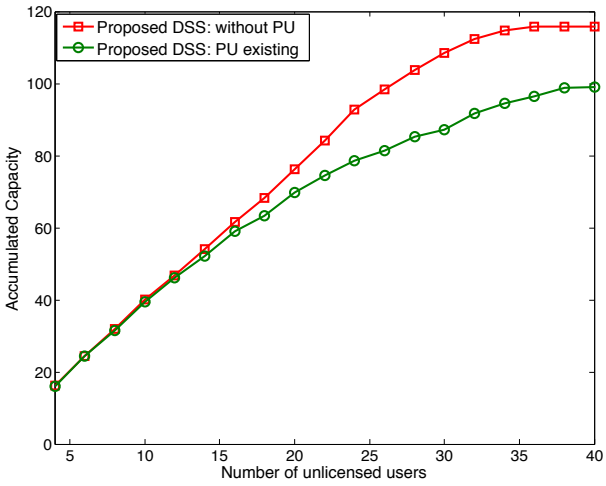

Fig. 8: Throughput comparison with the state of PUs

for both centralized algorithm and distributed algorithm. Two typical configurations on the predefined SINR threshold are considered, i.e. $\gamma_0 = 5$dB and $\gamma_0 = 10$dB.

The comparison results of centralized and distributed method are shown in Fig. 6. It is noted that the centralized scheme indeed has some advantage over a distributed one in terms of the accumulated capacity in various SINR requirements. This is easy to understand. A centralized algorithm possesses a control center to manage the information of all SUs and allocate the spectrum resource according to the complete information. In this case, all SUs are required not only to exchange information between their transmitter and receiver, but also to report their information to a control center, which, in turns, aggravates the coordination overhead. In other words, the centralized algorithm gains more system throughput, at the cost of higher resource consumptions and signaling overheads. In comparison, for a distributed algorithm, the information exchange only exists in the shared pair (i.e. two SUs). Thus, the resource demanding and signaling overhead can be reduced significantly. Despite the reduced system throughput, a distributed algorithm enables the practical balance between performance and cost.

2) The Main-Lobe Width: We then investigate the effects of the main-lobe width θ_n. In the simulations, the predefined SINR threshold is $\gamma_0 = 10$dB, and three configurations main-lobe width are considered, i.e., $\theta_n = \theta = 30°$, $\theta_n = \theta = 60°$, and $\theta_n = \theta = 90°$ respectively. The number of SUs in UDN will increase from 4 to 35.

We note from Fig. 7 that, in various main-lobe width settings, the network throughput of the new scheme will significantly outperform conventional methods. By providing the more precise spatial separation and the even reduced crosstalk, the number of accommodated SUs will be increased, whilst the mutual interference will be controlled effectively by the distributed learning scheme. Thus, the achievable maximum throughput will be increased as a main-lobe width deceases. Taking $\theta = 60°$ for example, its achievable throughput is roughly three times of the conventional method. Besides, the rationale behind the saturation effect in this figure is explained as follows. Under the given parameter settings, the number of maximum accommodated shared links exists, which is further demonstrated in the following C part. Similarly, a compromise between the achievable throughput and the implementation complexity should be made. As expected, a smaller main-lobe width requires more antennas and more complicated steering scheme.

3) The PUs' Status: In the above simulations, we assume there will be no active incumbent/PU. Once active incumbents are detected, the harmful interference to them should be avoided. In this case, it is understood that, due to the exclusive regions introduced by PUs, the performance spatial reuse will be compromised, and therefore, the network throughput of shared access will be reduced. In numerical analysis we assume there are 5 PUs, and each of them occupied one licensed channel. The predefined SINR threshold is $\gamma_0 = 10$dB, and the main-lobe width is $\theta_n = \theta = 30°$. Simulation results are plotted in Fig. 8. For the temporal-spatial reuse scheme, relying on the partial information of PU, SUs are allowed to share channels via properly adapted beams. Although there is performance degradation aroused by exclusive regions, the network throughput via shared access seems still to be appealing for UDN.

C. Maximum Accommodated Users

Besides the system throughput, the number of maximum accommodated SUs is another important metric for UDN. Therefore, we present the performance of maximum accommodated SUs in UDN scenario.

As noted by previous analysis, there exists a maximum number of accommodated SUs, given the main-lobe width and the SINR threshold. That is, if the accommodated SU is further increased, the network throughput will become saturated, while the link quality of each shared links will be deceased. In the last simulation, we further investigate the maximum number of accommodated SUs in the shared access. It should be noted that, for UDN application, the main purposes are to maximize the network aggregate throughput

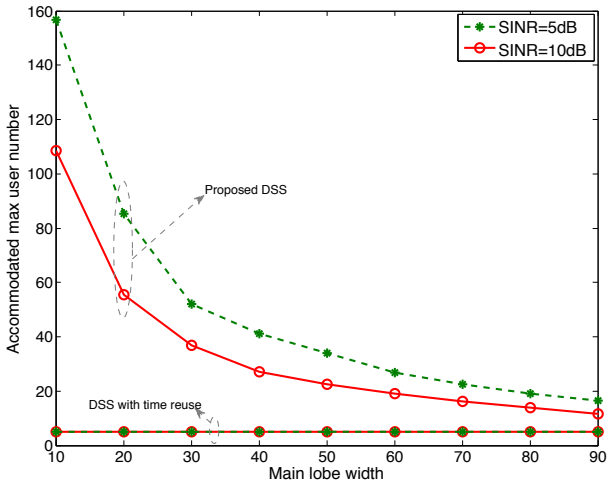

Fig. 9: Maximum accommodated SUs number

and accommodate as many as possible shared links, yet with SINR of each link higher than a tolerable threshold, rather than the achievable data rate of one single SUs. As shown by Fig. 9, the maximum number is associated with the main-lobe width as well as the predefined SINR threshold. From the numerical analysis, we find that: (i) the smaller of the main-lobe width and the SINR threshold, the more SUs will be accommodated; and (ii) the new access scheme will accommodate sufficiently larger number of users, by properly configuring the main-lobe width and the SINR threshold, while the conventional methods unfortunately will be inadequate in UDN scenarios.

VI. Conclusion

In this paper, dynamic spectrum access schemes for mm-Wave UDN have been studied. First, we established a new channel access model which enables the temporal-spatial reuse of spectrum, where flexible beams are assumed to provide the spatial separations and reduce the co-link interferences among SUs. Then, we formulated the channel access problem as one non-cooperative game, where the accumulated capacity of multiple SUs is served as its utility function, instead of a simple Kronecker delta function. Then, the existence of this pure NE is rigorously proved. A decentralized Q-learning algorithm with the self-adaption and low-complexity is proposed. A strict proof of the algorithm convergence and stability is also provided. Finally, it is demonstrated via numerical simulations that the fine-grained multi-dimensional shared access can significantly enhance the network throughput and the accommodated shared users with little access latency. Thus, our new shared access scheme will be of great promise to the emerging UDN. Future works will focus on the exclusion of some partial information in the distributed access.

Acknowledgment

This work was supported by Natural Science Foundation of China (NSFC) under Grant 61471061 and Grant 61571100.

References

[1] P. Pirinen, "A brief overview of 5g research activities," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, 2014, pp. 17–22.

[2] X. Ge, S. Tu, G. Mao, C.-X. Wang, and T. Han, "5g ultra-dense cellular networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, 2016.

[3] S. Stefanatos and A. Alexiou, "Access point density and bandwidth partitioning in ultra dense wireless networks," *IEEE transactions on communications*, vol. 62, no. 9, pp. 3376–3384, 2014.

[4] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 gbps/ue in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2078–2101, 2015.

[5] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey."

[6] M. Ding, D. López-Pérez, G. Mao, P. Wang, and Z. Lin, "Will the area spectral efficiency monotonically grow as small cells go dense?" in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–7.

[7] A. K. Gupta, X. Zhang, and J. G. Andrews, "Sinr and throughput scaling in ultradense urban cellular networks," *IEEE Wireless Communications Letters*, vol. 4, no. 6, pp. 605–608, 2015.

[8] H. Claussen, I. Ashraf, and L. T. Ho, "Dynamic idle mode procedures for femtocells," *Bell Labs Technical Journal*, vol. 15, no. 2, pp. 95–116, 2010.

[9] C. Li, J. Zhang, M. Haenggi, and K. B. Letaief, "User-centric intercell interference nulling for downlink small cell networks," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1419–1431, 2015.

[10] A. Gupta and R. K. Jha, "A survey of 5g network: architecture and emerging technologies," *IEEE access*, vol. 3, pp. 1206–1232, 2015.

[11] F.-H. Tseng, H.-c. Chao, J. Wang *et al.*, "Ultra-dense small cell planning using cognitive radio network toward 5g," *IEEE Wireless Communications*, vol. 22, no. 6, pp. 76–83, 2015.

[12] B. Li, C. Zhao, M. Sun, and Z. Zhou, "Spectrum sensing for cognitive radios in time-variant flat-fading channels: A joint estimation approach," *Communications IEEE Transactions on*, vol. 62, no. 8, pp. 2665–2680, 2014.

[13] S. Haykin and P. Setoodeh, "Cognitive radio networks: the spectrum supply chain paradigm," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 1, pp. 3–28, 2015.

[14] A. A. El-Sherif and K. R. Liu, "Joint design of spectrum sensing and channel access in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 6, pp. 1743–1753, 2011.

[15] C. Pan, J. Wang, W. Zhang, B. Du, and M. Chen, "Power minimization in multi-band multi-antenna cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5056–5069, 2014.

[16] D. N. Nguyen and M. Krunz, "Power minimization in mimo cognitive networks using beamforming games," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 916–925, 2013.

[17] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wang, and T. Q. Quek, "Resource allocation for cognitive small cell networks: A cooperative bargaining game theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3481–3493, 2015.

[18] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 2, pp. 180–194, 2012.

[19] D. Niyato and E. Hossain, "Competitive spectrum sharing in cognitive radio networks: a dynamic game approach," *IEEE Transactions on wireless communications*, vol. 7, no. 7, pp. 2651–2660, 2008.

[20] J.-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei, "Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3251–3271, 2010.

[21] J. Lundén, M. Motani, and H. V. Poor, "Distributed algorithms for sharing spectrum sensing information in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4667–4678, 2015.

[22] D. Wu, L. Zhang, and J. Zhou, "Self-organized spectrum access in small cell networks: A noncooperation interference minimization game solution," in *Wireless Communications & Signal Processing (WCSP), 2015 International Conference on*. IEEE, 2015, pp. 1–5.

[23] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic

stochastic learning solution," *IEEE transactions on wireless communications*, vol. 11, no. 4, pp. 1380–1391, 2012.
[24] E. G. Larsson and E. A. Jorswieck, "Competition versus cooperation on the miso interference channel," *IEEE Journal on selected areas in Communications*, vol. 26, no. 7, pp. 1059–1069, 2008.
[25] Z. Juntao, F. Wei, Z. Ming, and W. Jing, "Coordinated multi-user spectrum sharing in distributed antenna-based cognitive radio systems," *China Communications*, vol. 13, no. 1, pp. 57–67, 2016.
[26] B. Li, S. Li, A. Nallanathan, and C. Zhao, "Deep sensing for future spectrum and location awareness 5g communications," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 7, pp. 1331–1344, 2015.
[27] B. Li, S. Li, A. Nallanathan, Y. Nan, C. Zhao, and Z. Zhou, "Deep sensing for next-generation dynamic spectrum sharing: more than detecting the occupancy state of primary spectrum," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2442–2457, 2015.
[28] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.
[29] B. Li, Z. Zhou, W. Zou, X. Sun, and G. Du, "On the efficient beamforming training for 60ghz wireless personal area networks," *Wireless Communications IEEE Transactions on*, vol. 12, no. 2, pp. 504–515, 2013.
[30] B. Li, Z. Zhou, H. Zhang, and A. Nallanathan, "Efficient beamforming training for 60-ghz millimeter-wave communications: a novel numerical optimization framework," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 703–717, 2014.
[31] R. De Francisco and D. T. Slock, "An optimized unitary beamforming technique for mimo broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 9, no. 3, pp. 990–1000, 2010.
[32] A. Tajer, N. Prasad, and X. Wang, "Beamforming and rate allocation in miso cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 362–377, 2010.
[33] "Channel models for 60 ghz wlan systems," *IEEE 802.11-0910334*, 2010.
[34] J. Park, H. Lee, and S. Lee, "Mimo broadcast channels based on sinr feedback using a non-orthogonal beamforming matrix," *IEEE Transactions on Communications*, vol. 60, no. 9, pp. 2534–2545, 2012.
[35] A. I. Sulyman, A. T. Nassar, M. K. Samimi, G. R. Maccartney, T. S. Rappaport, and A. Alsanie, "Radio propagation path loss models for 5g cellular networks in the 28 ghz and 38 ghz millimeter-wave bands," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 78–86, 2014.
[36] A. I. Sulyman, A. Alwarafy, and G. R. MacCartney Jr, "Directional radio propagation path loss models for millimeter-wave wireless networks in the 28-, 60-, and 73-ghz bands," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, p. 6939, 2016.
[37] Y. Xiao, G. Bi, and D. Niyato, "Game theoretic analysis for spectrum sharing with multi-hop relaying," *IEEE Transactions on Wireless Communications*, vol. 10, no. 5, pp. 1527–1537, 2011.
[38] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Decentralized dynamic spectrum access for cognitive radios: cooperative design of a non-cooperative game," *IEEE Transactions on Communications*, vol. 57, no. 2, pp. 459–469, 2009.
[39] G. Bacci, S. Lasaulce, W. Saad, and L. Sanguinetti, "Game theory for networks: A tutorial on game-theoretic tools for emerging signal processing applications," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 94–119, 2016.
[40] M. Harmon and S. Harmon, "Reinforcement learning: a tutorial.{online}," 2000.
[41] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.
[42] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.
[43] H. Li, "Multi-agent q-learning for competitive spectrum access in cognitive radio systems," in *Networking Technologies for Software Defined Radio (SDR) Networks, 2010 Fifth IEEE Workshop on*. IEEE, 2010, pp. 1–6.
[44] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.